DOCUMENT RESUME

ED 099 417 TM 004 081

AUTHOR	Waller, Michael I.
TITLE	Removing the Effects of Random Guessing from Latent Trait Ability Estimates. Research Bulletin No. 74-32.
INSTITUTION	Educational Testing Service, Princeton, N.J.
REPORT NO	ETS-RB-74-32
PUB DATE	Aug 74
NOTE	52p.; Paper presented at the Annual Meeting of the American Educational Research Association (Chicago, Illinois, April 1974)
EDRS PRICE	MF-$0.75 HC-$3.15 PLUS POSTAGE
DESCRIPTORS	Goodness of Fit; *Guessing (Tests); Individual Differences; Item Analysis; *Multiple Choice Tests; Response Mode; *Response Style (Tests); *Statistical Analysis; *Testing Problems
IDENTIFIERS	Latent Trait Model

ABSTRACT

In latent trait models the standard procedure for handling the problem caused by guessing on multiple choice tests is to estimate a parameter which is intended to measure the "guessingness" inherent in an item. Birnbaum's three parameter model, which handles guessing in this manner, ignores individual differences in guessing tendency. This paper presents a model or procedure which uses the information contained in the interaction between a person and an item to remove the effects of random guessing from estimates of ability, difficulty, and discrimination. Simulated and real data are presented which support the model in terms of fit and information. (Author/RC)

ED 099417

RESEARCH BULLETIN

RB-74-32

REMOVING THE EFFECTS OF RANDOM GUESSING FROM LATENT TRAIT ABILITY ESTIMATES

Michael I. Waller
The University of Chicago

U S DEPARTMENT OF HEALTH
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

This paper was presented at the Annual Convention of the American Educational Research Association, Chicago, Illinois, April 19, 1974.

Educational Testing Service
Princeton, New Jersey
August 1974

TM 004 081

Removing the Effects of Random Guessing
From Latent Trait Ability Estimates

Abstract

In latent trait models the standard procedure for handling the problem caused by guessing on multiple choice tests is to estimate a parameter which is intended to measure the "guessingness" inherent in an item. Birnbaum's three parameter model, which handles guessing in this manner, ignores individual differences in guessing tendency. This paper presents a model or procedure which uses the information contained in the interaction between a person and an item to remove the effects of random guessing from estimates of ability, difficulty, and discrimination. Simulated and real data are presented which support the model in terms of fit and information.

Removing the Effects of Random Guessing From Latent Trait Ability Estimates

Michael I. Waller[1]

The University of Chicago

1. Introduction

It is well known that individuals vary in their tendency to guess randomly on multiple choice tests. With latent trait models the standard procedure for handling random guessing on multiple choice tests is to estimate a parameter which is intended to represent the "guessingness" inherent in an item (see, e.g., Birnbaum, 1968). Such a three parameter or item-guessing model ignores individual variation in guessing tendency. Within classical test theory the "correction for guessing" (see, e.g., Diamond and Evans, 1973) also estimates guessingness, although in this case the estimate is a function of the number of wrong responses made by an individual.

We argue here, that with models designed to estimate ability, there is no need to estimate random guessing behavior and correct for it, whether such behavior is attached to the item or the person. In either case our primary interest is in estimating ability. The models are intended for that purpose, and our interest in guessing arises only from an interest in eliminating the "noise" it creates in ability estimation. Accordingly, consideration of the problem in terms of eliminating the noise rather than estimating guessing and correcting for it should be more fruitful, and this is the view taken here. Since a large proportion of guesses occur when low ability subjects meet items which are too difficult for them,

---

[1]Present address: Educational Testing Service, Princeton, N.J.

Panchapakesan (1969) has suggested omitting low ability subjects entirely when estimating the item parameters. However, these subjects can contribute relevant information concerning easy items. The procedure presented here represents an improvement over her idea in two important ways. First, the information contributed by every subject is used during calibration of the instrument; but is used at only those places where one may be reasonably sure it is valid information. Second, the procedure yields a criterion for measuring the adequacy of this method in accounting for random guessing.

In the present paper we propose a latent trait model or procedure which uses the information contained in the interaction between a person and an item to remove most of the effects of random guessing from estimates of ability (and from estimates of both item parameters, difficulty and discrimination). This is accomplished through a modification of the free response model removing those item-person interactions characterized by the item being too difficult for the person and therefore likely to invite guessing. The basic assumptions of latent trait models, unidimensionality and local independence, are also made here.

The statistical procedures we derive for the model include: 1) estimation of the item parameters; 2) estimation of latent ability and measurement error; 3) an item-by-item test of goodness of fit of the model; and 4) an evaluation of the information recovered by a test. The model is equally applicable to the normal or logistic response laws. Although the discussion in this paper is in terms of binary scored items, the model is immediately

generalizable to the nominal category scoring model (Bock, 1972), as well as the graded response model (Samejima, 1972).

2. The Data

Suppose that each of N subjects respond to n multiple choice items, each item containing A_j alternatives, $j=1,\ldots,n$. The response of the ith subject to the jth item may be thought of as right or wrong. Omitted items are treated as wrong responses. While this treatment of omits is considered a flaw in the three parameter model (Lord,1968, p. 992), we feel the present model which considers each item-person interaction separately is better able to justify such treatment (see section 3).

3. The Response Mode

Let θ_i be a value on the continuum of latent ability underlying the responses to the test items, and let the event that a subject of ability θ_i responds correctly to item j be denoted $r_{ij}=1$. Then the free response model can be represented by equation (1).

$$(1) \qquad \Pr(r=1|\theta_i)=P_{ij} = F(Y_{ij}) \; ;$$

$$\text{where:} \qquad F(Y_{ij})=\int_{-\infty}^{Y_{ij}} \frac{1}{\sqrt{2\pi}} e^{-t^2/2}\, dt$$

$$\text{or:} \qquad F(Y_{ij})=\exp(Y_{ij})/(1+\exp(Y_{ij}))$$

$$\text{in which} \qquad Y_{ij}=b_j + a_j\theta_i .$$

The quantities b_j and a_j are the item parameters, difficulty and discrimination respectively, associated with item j.

To obtain estimates of ability removing random guessing, or A.R.R.G. estimates, the required adjustment to the free response model is represented by equation 2.

$$(2) \qquad P_{ij} = \begin{cases} F(a_j, b_j|\theta_i) \; ; \; F \geq P_c \\ g_{ij} \qquad\qquad ; \; F < P_c \end{cases}$$

where g_{ij} = Pr (Person i guesses item j correctly given $F \leq P_c$), and where P_c is some small probability. What use might be made of the set of items where $F \leq P_c$ is considered elsewhere (Waller, 1974). For our purpose consider what effect this procedure has on estimation of the principal parameters of the model, b_j, a_j and θ_i.

The basic idea is to base the estimate of any person's ability on only those items for which there is a reasonable chance that the person achieved the correct response through the interaction of his ability and the item characteristics. That is, an item which is very difficult for a particular person is an item which invites guessing and therefore is eliminated from consideration in estimating the person's ability (also that person's response is removed from the sample used to calibrate such an item). Whether or not the person guesses on such an item has no substantial effect on the estimate of his ability, because these item-person interactions are removed from the estimation procedure.

More specifically, we obtain a preliminary estimate of a subject's ability from the approximate transformation of his per cent correct, inverse normal or logistic. This gives us a rough idea of where the subject belongs

on the ability continuum. In each iteration of the estimation procedure the probability of a correct response, F, is estimated for each subject's response to each item. The A.R.R.G. procedure simply omits from estimation any interaction for which this estimated probability is less than some small probability, the cutoff point, P_c. An adequate method for determining the appropriate value of P_c is readily available when testing the fit of the model.

The method presented here treats omits the same as wrong responses. In support of this treatment it may be argued that there exists a probability, P_c, which can be used to divide all responses into two approximate groups, those responses which are made solely on the basis of the subjects' ability, and those responses which for some examinees represent random guessing and therefore as a group contribute more noise than information an estimation procedure. The first group consists of responses which occur when the probability of a correct response, P_{ij}, is above the point P_c. It is assumed that subjects in this group either know the correct answer or do not, and if not, either omit the question or respond incorrectly due

to misinformation. Support for such a model of behavior comes from a recent study by Bock (1972) which shows that the selection of certain wrong alternatives is representative of positive ability. The subject knows enough to choose what he believes to be the correct alternative, but not enough to make the finer discrimination which would enable him to choose what is in fact the correct alternative. If he does not omit the item, his partial knowledge misinforms him and leads him to select an incorrect alternative.

The second group consists of responses in which P_{ij} is less than P_c; and across a sample of examinees two behaviors are assumed to occur. Non-guessers (or low risk takers) continue to behave as all subjects behave above P_c, they either know the correct answer or do not, and if not, either omit or respond incorrectly. Guessers (or high risk takers) either know the answer or do not; but if not, these subjects will tend to guess, in which case, they will be correct $100/A_j$% of the time, where A_j is the number of alternatives. It is assumed that the procedure is robust with respect to minor differences in the point at which individuals may begin to guess randomly.

4. Estimation

There are a number of methods for obtaining estimates of the parameters of latent trait models (see e.g. Birnbaum, 1968, Bock, 1972, Bock and Lieberman, 1970, Lord, 1968). The method described in the present study may be termed conditional estimation (Bock, 1972). This method as applied to the three parameter item-guessing model is described in Kolakowski and Bock (1970). As the A.R.R.G. procedure for removing the effects of random guessing from latent trait parameter estimates is a modification of the free response model, we first review the estimation procedure for the free response model from which estimation removing random guessing is easily seen[2]. The estimation procedure is outlined in terms of a general response relation

$$P_{ij} = F(Y_{ij})$$

where F may be any monotonic function which maps the real line into the unit interval, e.g. the normal ogive or logistic ogive.

---

[2] Also, when the item-guessing parameter, g_j, in the three parameter item-guessing model, $P_{ij} = g_j + (1-g_j) F_{ij}$, is treated as a constant during maximum likelihood estimation, the procedure described here is immediately generalized to the item-guessing model.

Maximum likelihood estimation of ability

Suppose we have estimates of the item parameters, discrimination and difficulty, of n dichotomous items. These estimates might be values based on a previous calibration of the testing instrument or estimates from the previous cycle during calibration of the instrument. For the i^{th} person's encounter with the j^{th} item,

let $r_{ij} = 1$ denote a correct response,

and $r_{ij} = 0$ denote an incorrect response.

Further, let $\underline{r}_i' = \{r_{i1}, r_{i2}, \ldots, r_{in}\}$ denote the response vector of person i. Thus, under the assumption of local independence--that responses of subjects with the same ability to different items are statistically independent--equation (3),

$$(3) \qquad L_i = Pr(\underline{r}_i) = \prod_{j=1}^{n} P_{ij}^{r_{ij}} Q_{ij}^{(1-r_{ij})},$$

is the joint probability function of the response vector for person i. To obtain the maximum likelihood (m.l.) estimate of θ_i we obtain the first and second partial derivatives of the log of equation (3). These equations (omitting the subscripts) are:

$$(4) \qquad \ell_\theta = \frac{\partial \ell}{\partial \theta} = \sum_{j=1}^{n} \left[\frac{r - P}{P Q}\right] \frac{\partial P}{\partial \theta}, \text{ and}$$

$$(5) \qquad \ell_{\theta\theta} = \frac{\partial^2 \ell}{\partial \theta^2} = \Sigma \left\{ \left[\frac{r-P}{PQ} \cdot \frac{\partial^2 P}{\partial \theta^2} \right] - \left(\frac{\partial P}{\partial \theta} \right)^2 \cdot \left[\frac{PQ + (r-P)(Q-P)}{P^2 Q^2} \right] \right\} .$$

Considering equation (4), ℓ_θ, we see that the equation $\ell_\theta = 0$ is not easily solved for an explicit statement of $\hat{\theta}_i$. However, the solution is available by means of an iterative process. For example, Newton-Rapheson iteration allows us to obtain a maximum likelihood estimate of θ_i. For one variable the estimate of this parameter at the $(k+1)^{st}$ iteration, $\hat{\theta}_i^{k+1}$, is given by equation (6).

$$(6) \qquad \hat{\theta}^{k+1} = \hat{\theta}^{k} - \ell_\theta / \ell_{\theta\theta}$$

with the two partials being evaluated at the previous estimate of $\hat{\theta}$, $\hat{\theta}^k$. This procedure is repeated until the correction, $\ell_\theta / \ell_{\theta\theta} = \overline{\Delta\theta}$ is less than some previously specified criterion, say .001.

Conditional estimation of item parameters by maximum likelihood

Conditional estimation of the item parameters, that is, the calibration of the instrument, also uses previously obtained estimates of the parameters not in question, in this case abilities. However, the time required to estimate

the item parameters can be greatly reduced if the data are reassembled in a binomial form, and the principle of local independence relaxed somewhat. It is found to be expedient to assume that "subjects whose latent ability is in the 'neighborhood' of θ respond independently to different items. The purpose of this relaxation is to justify grouping subjects for whom provisional estimates of latent scores are similar. It is assumed that the actual latent scores of subjects in such groups are confined to a sufficiently small neighborhood to assure independent responses. The question of how small this neighborhood should be to justify the local independence assumption is left to further empirical study" (Bock, 1972, p. 37).

Under the relaxed assumption of local independence, we can order the subjects by ability and divide them into q fractiles. The number of subjects per fractile, N_i, may be assumed to follow a specific distribution, for example N(0,1), or the so called "empirical" assumption can be made that there are an equal number of subjects in each fractile (Kolakowski and Bock, 1970, p. 5). In either case let s_{ij} = the number of subjects in the i^{th} fractile who got item j correct, and let $\underline{s}' = [s_{1j}, \ldots, s_{qj}]$ so that $\underline{s}'$ represents the vector of the item responses across the q fractiles. Then under the relaxed

assumption of local independence, equation (7)

$$(7) \qquad L_j = Pr\ (\underline{s}) = \prod_{i=1}^{q} \frac{N_i!}{s_{ij}!\ (N_i - s_{ij})!} P_{ij}^{s_{ij}} Q_{ij}^{(N_i - s_{ij})},$$

is the joint probability function of the response vector for item j. As above, P_{ij} is the response relation and is a function of the item parameters and the θ_i associated with the i^{th} fractile[3]. The value to be used for θ_i in the estimation of the item parameters depends on the assumed distribution of abilities: If a normal distribution is assumed for θ, the normal deviate corresponding to the centroid of each fractile is used; if an empirical assumption is made, the value of θ used is the median value in the fractile (Kolakowski and Bock, 1970, p.5).

As in the case of ability estimates, we will use as our estimates of item parameters, a_j and b_j, those values

---

[3] The θ_i are standardized to a mean of zero and variance of one.

which maximize $\ell = \log L_j$. In order to obtain joint estimates of these parameters, we obtain all first and second partial derivatives of ℓ with respect to the item parameters. These equations (omitting the subscripts) are:

$$(8) \qquad \ell_b = \frac{\partial \ell}{\partial b} = \sum_i^q N_i \left(\frac{p-P}{PQ}\right) \cdot \frac{\partial P}{\partial b} ,$$

where $p_{ij} = s_{ij}/N_i$

$$(9) \qquad \ell_a = \frac{\partial \ell}{\partial a} \sum_i^q N_i \left(\frac{p-P}{PQ}\right) \cdot \frac{\partial P}{\partial a} ,$$

$$(10) \qquad \ell_{bb} = \frac{\partial^2 \ell}{\partial b^2} \sum_i^q N_i \left\{ \left(\frac{p-P}{PQ}\right) \cdot \frac{\partial^2 P}{\partial b^2} - \left(\frac{\partial P}{\partial b}\right)^2 \cdot \left[\frac{PQ + (p-P)(Q-P)}{P^2Q^2}\right] \right\} ,$$

$$(11) \qquad \ell_{ba} = \frac{\partial^2 \ell}{\partial b \partial a} = \sum_i^q N_i \left\{ \left(\frac{p-P}{PQ}\right) \cdot \frac{\partial^2 P}{\partial b \partial a} - \left(\frac{\partial P}{\partial b}\right) \cdot \left(\frac{\partial P}{\partial a}\right) \cdot \left[\frac{PQ + (p-P)(Q-P)}{P^2Q^2}\right] \right\} ,$$

$$(12) \qquad \ell_{aa} = \frac{\partial^2 \ell}{\partial a^2} = \sum_i^q N_i \left\{ \left(\frac{p-P}{PQ}\right) \cdot \frac{\partial^2 P}{\partial a^2} - \left(\frac{\partial P}{\partial a}\right)^2 \cdot \left[\frac{PQ + (p - P)(Q - P)}{P^2 Q^2}\right] \right\},$$

Again we find that the equations of the first partials, $\ell_b = 0$ and $\ell_a = 0$ are not easily solved in closed form and we again rely on a Newton-Rapheson procedure. This is accomplished for the case of two variables by writing out the first two terms of the Taylor expansion of these two equations as in equations (13) and solving the resulting system of linear equations in the corrections, Δb_j and Δa_j. These corrections are added to the k^{th} stage estimates, b_j^k and a_j^k, to form the $(k+1)^{st}$ stage estimates. (Hildebrand, 1956, pp. 443-51.) Omitting the j subscripts we have

$$(13) \qquad \begin{aligned} 0 &= \ell_b (a^k, b^k) + \ell_{bb} (a^k, b^k) \overline{\Delta b} + \ell_{ba} (a^k, b^k) \overline{\Delta a} \\ 0 &= \ell_a (a^k, b^k) + \ell_{ab} (a^k, b^k) \overline{\Delta b} + \ell_{aa} (a^k, b^k) \overline{\Delta a} \end{aligned}$$

where: $\overline{\Delta b} = b + \delta b$ and $\overline{\Delta a} = a + \delta a$.

As with ability estimates, this process is repeated until the corrections, $\overline{\Delta b}$ and $\overline{\Delta a}$, are both less than some criterion. These equations can be restated more compactly in matrix terms,[4]

$$(14) \qquad \begin{bmatrix} -\ell_b \\ \\ -\ell_a \end{bmatrix} = \begin{bmatrix} \ell_{bb} & \ell_{ba} \\ \\ \ell_{ba} & \ell_{aa} \end{bmatrix} \cdot \begin{bmatrix} \overline{\Delta b} \\ \\ \overline{\Delta a} \end{bmatrix}$$

or $-\ell = H \cdot \underline{\Delta}$

Equation (14) yields the following corrections:

$$\overline{\Delta b} = (-\ell_{aa}\ell_b + \ell_{ba}\ell_a)/\delta$$

$$\overline{\Delta a} = (\ell_{ba}\ell_b - \ell_{bb}\ell_a)/\delta$$

which are added to the k^{th} iteration estimates.

Here, $\delta = \text{Det}(H) = \ell_{aa}\ell_{bb} - \ell_{ab}^2$.

The Mathematics of the A.R.R.G. Estimation Procedure

With the estimation procedure of the free response model firmly in hand, the adjustment implied for estimation

---

[4] $\ell_{ab} = \ell_{ba}$.

of abilities with the A.R.R.G. procedure is simple and straightforward. To understand the implications of equation (2) in terms of its effect on the m.l. estimation procedure outlined above we need only consider the first and second partial derivatives of P_{ij} with respect to θ_i. Observe, the A.R.R.G. model implies the following:

$$P_{\theta} = \frac{\partial P_{ij}}{\partial \theta_i} = \begin{cases} \frac{\partial F(Y_{ij})}{\partial \theta_i} & F(Y_{ij}) \geq P_c \\ 0 & F(Y_{ij}) < P_c \end{cases}$$

(15)

$$P_{\theta\theta} = \frac{\partial^2 P_{ij}}{\partial \theta_i^2} = \begin{cases} \frac{\partial^2 F(Y_{ij})}{\partial \theta_i^2} & F(Y_{ij}) \geq P_c \\ 0 & F(Y_{ij}) < P_c \end{cases}$$

As expected, one or the other of these derivatives are multipliers in the expressions for ℓ_{θ} and $\ell_{\theta\theta}$ given in equations (4) and (5), consequently, the response to any item for which $F(Y_{ij})$ is less

than P_c will not affect the estimation of the ability being considered.

The A.R.R.G. Procedure:
Estimation of Item Parameters

As in the case of ability estimates, the modification of the procedure used to obtain estimates of the difficulty and discrimination, b_j and a_j, under the free response model is most readily seen by considering the first and second partials of P_{ij} with respect to these parameters. The form of these equations is identical to that given with respect to ability in equation (15). In effect we are again assuming that those item-subject interactions which produce provisional estimates of P_{ij} which are deemed unreasonable will not produce relevant information for estimation of the item parameters; and consequently, the derivatives associated with such interaction are zero.

Measurement Error

The estimate of the asymptotic variance of the m.l. estimator, $\hat{\theta}$, is obtained from Fisher's information function which can be stated:

$$I(\theta) = -E\left[\frac{\partial^2 (\ell = \log L)}{\partial \theta^2}\right] ,$$

where E indicates expectation.

It can be shown that asymptotically, $\hat{\theta}$ has a normal distribution with mean θ and variance $1/I(\theta)$; i.e.,

$$\hat{\theta} \sim N(\theta, 1/I(\theta)) .$$

Clearly, the larger the value of $I(\theta)$, the information, the more precise will be our estimate of ability. We will use the information contained in the ability estimates resulting from different models to make comparisons of the models.

While we estimate the precision of every ability estimate, for purposes of general comparison we would like to obtain a statement of the information contained in any test concerning a general level of ability. In other words, subjects of identical ability should respond stochastically the same to a given item. At expectation the differences between people vanish and we are able to obtain a statement for the information contained in a test at a general ability level by observing that at expectation equation (5), the second partial of ℓ with res-

pect to θ, is simplified in that terms which contain r-P vanish (see e.g. Birnbaum, 1968). Therefore the statement of the information contained in a test at a general level of ability is:

$$I(\theta) = \ell_{\theta\theta} = \sum_{j}^{n} \frac{\left(\frac{\partial P}{\partial \theta}\right)^2}{PQ} \quad .$$

As has been shown, for each individual item, a test of deviation from the model can be obtained since Q_j, equation (16), is distributed as a Pearsonian χ^2 on $q - 2$ degrees of freedom (Bock and Jones, 1968, pp. 51-60). Finally a test of fit for the test as whole, χ^2_{Test}, is obtained by summing over items the residual sums of squares, Q_j, and comparing that sum to a χ^2 on $f = \sum_{j=1}^{n} (q-2)-2 = n(q-2)-2$ degrees of freedom.

$$(16) \qquad Q_j = \sum_{i=1}^{q} \frac{N_i (p_{ij} - P_{ij})^2}{P_{ij} Q_{ij}}$$

$$(17) \qquad \chi^2_{Test}(f) = \sum_{j=1}^{n} Q_j \quad .$$

The test of fit enables us to identify the best cutoff point to use in applying this procedure. If we ignore random guessing responses when they are in fact present in the data, i.e., fit a free response model to data contaminated by random guessing, we would expect the resulting fit to be poorer than the fit resulting from a model which adequately accounts for random guessing responses. Within the present context if we allow all the responses to remain in the estimation procedure the fit will be poorer than if we omit those responses which may result from random guessing: Too many people at lower ability levels will get the more difficult items correct. On the other hand, if we remove too many responses from the estimation procedure, two few subjects will appear to be getting the more difficult items correct and we again will observe a poorer fit. Consequently by beginning with a cutoff point of $P_c = 0$ (i.e., a free response analysis) and increase the cutoff point we should observe an improved fit up to a point followed by a poorer fit. The cutoff point which produces the best fit is the proper value for P_c.

The effect on the recovered item parameters of such a procedure is straightforward. Underestimating the

cutoff, retaining responses at a level where some responses are random guesses, results in underestimating the difficulty and discriminating power of the effected items. Overestimating the cutoff point results in an overestimation of the effected item's difficulty and discrimination.[5]

Note that fit occurs with respect to items. Given a set of items already calibrated, the effect on ability estimation of removing more items than necessary to remove the effects of random guessing is simply less information and consequently less precise ability estimates.

---

[5] In this formu'ation, large negative values of the difficulty parameter correspond to difficult items; consequently, over (under)-estimation as used here refers to the absolute value of the difficulty parameter.

5.1 Simulated Data

In this section we examine a number of analyses of two kinds of simulated data: Data sets simulating free response behavior or Non-guessing data sets; and data sets simulating random guessing of the kind modeled by equation (2), i.e., Guessing data sets. A non-guesser's response vector is generated by assuming values for his ability parameter and for all item parameters, calculating the true probability of a correct response for each item-person interaction, and then comparing this probability to a random number between zero and one. A guesser's response vector is generated in the same manner with the exception that for those item-person interactions in which this calculated probability is less than the cutoff point, e.g., $P_c = .05$, each subject is assumed to guess in an essentially random manner. The same random sequence and the same set of abilities are used for both guessers and non-guessers, so that the response vectors of guessers and non-guessers differ only on the subset of items where the calculated probability of a correct response is less than the cutoff point. With 5 as the number of alternatives, a guesser will receive a correct response in a random manner in 20% of this subset of items, whereas a non-guesser will receive a correct response on less than P_c% of such items.

A Non-guessing data set is composed entirely of non-guessing subjects; whereas a Guessing data set contains approximately twenty-five per cent guessing subjects.

Two pairs of data sets, each composed of one Non-guessing and one Guessing data set, were generated. Both sets in a pair utilized the same assumed item parameters. The first pair used a more or less idealized set of item parameters with difficulties from -2.2 to 2.2 in steps of .1 and constant discriminations. The second pair used item parameters obtained from a previously calibrated instrument with a similar range of difficulties, but with widely varying discriminations.

We present only the characteristics of analyses of the two sets of simulated data with constant discriminations, a set of free response or Non-guessing data and a set of Guessing data with a true P_c equal to .05.[6] When variation in discriminating power is introduced into data contaminated by guessing the results presented below resulting from the constant discrimination data are replicated with one exception. Varying discriminations introduce a component of variance into the procedure which

---

[6] All analyses were performed assuming a normal distribution of abilities, and the Normal Ogive Response Relation.

is not completely accounted for by the estimated measurement error. Out of 480 subjects we would expect 95% confidence limits to cover all but 24 of the true abilities. Forty-five abilities were in fact missed by their estimated confidence limits calculated from $P_c = .05$ analysis of the varying discrimination data set.

Sets of Guessing data were also generated simulating individual cutoff points of .10 and .20 with no significant changes in the results described in this paper; particularly, the best fit always occurred with $P_c = .05$. Since the test of fit is made with respect to the item parameters, increasing the point at which guessing begins for any individual does not produce much of an effect on this aspect of the estimation procedure: when the subjects are grouped for item parameter estimation, the proportion of correct <u>guesses</u> in most fractiles remains the same. Observe that the proportion of correct guesses is the product of the proportion of guessers and the probability of a correct response; i.e., with 25% of the sample simulating guessers and a 5 choice test the proportion of correct guesses on any item which admits guessing is .05.

Consequently, increasing the proportion of guessers in the sample will result in an increase in the proportion of correct guesses on each item in every fractile affected by guessing; consequently such an increase should affect the estimation procedure. To demonstrate the behavior of the model in this respect, sets of Guessing data were analyzed in which the proportion of guessers was set at 50% and in this case the best analysis occurred with P_c at the implied level of correct guesses, .50 x 1/5 or .10. The implications of this for analyses of real data is that identification of the best cutoff point reflects the proportion of guessers in the sample and not the probability at which any individual begins to guess.

Table 1 gives the values of statistics for the sample of abilities used to generate both sets of data, and the values of these parameters as calculated from the recovered sets of abilities from the analyses of the Non-guessing data with constant discriminations utilizing different cutoff points. Table 2 presents the same values for analyses of the Guessing data. Figures (1 to 3) contain plots of the 45 pairs of recovered item parameters from the three analyses of the simulated Guessing data (A= Discriminations, B = Difficulty).

TABLE 1
MOMENTS OF THE TRUE AND RECOVERED SAMPLES OF ABILITIES
NON-GUESSING DATA (CONSTANT DISCRIMINATIONS)

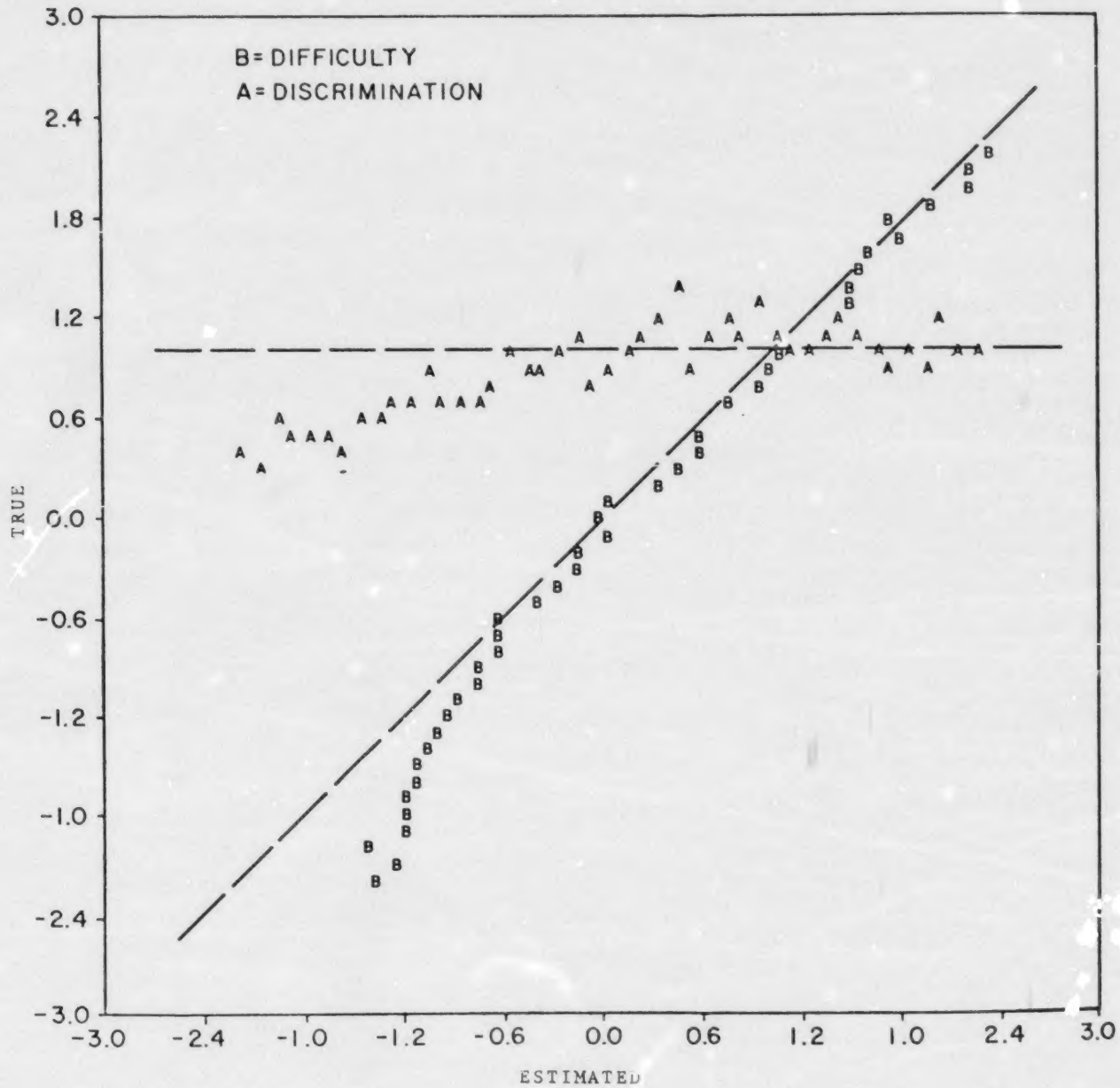
N = 478	True	P_c = .00 (Free Response)	P_c = .05
Mean	-0.008	-0.003	0.012
(S.E.)		(0.046	
Variance	0.964	1.007	0.979
(95% C.L.)		(0.89, 1.15)	(0.87, 1.12)
Skewness	-0.201	-0.090	-0.146
(S.E.)		(0.112)	
Kurtosis	-0.380	-0.292	-0.373
(S.E.)		(0.223)	
		Range Statistics	
Minimum Ability	-2.66	-2.71	-2.63
Maximum Ability	+2.63	+2.70	+2.49
Range	5.29	5.41	5.12
χ^2_{Test} on 358 d.f.		258.48	311.46

TABLE 2
MOMENTS OF THE TRUE AND RECOVERED SAMPLES OF ABILITIES
GUESSING DATA (CONSTANT DISCRIMINATIONS)

N = 478	True	Analysis P_c = .00	P_c = 0.05	P_c = 0.10
Mean	-.008	+0.020	+0.010	+.026
(S.E.)			(0.046)	
Variance	.964	1.089	0.975	0.971
(95% C.L.)		(0.96, [b]1.24	(0.86, 1.12)	(0.75, 1.11)
Skewness	-.201	+0.334[b]	-0.083	-0.193
(S.E.)			(0.112)	
Kurtosis	-.380	+0.431[a]	-0.315	-0.327
(S.E.)			(0.223)	
		Range Statistics		
Minimum Ability	-2.66	-2.77	-2.60	-2.97
Maximum Ability	+2.63	+3.90	+2.68	+2.37
Range	5.29	6.67	5.28	5.34
χ^2_{Test} on 358 d.f.		371.25	271.01	399.89

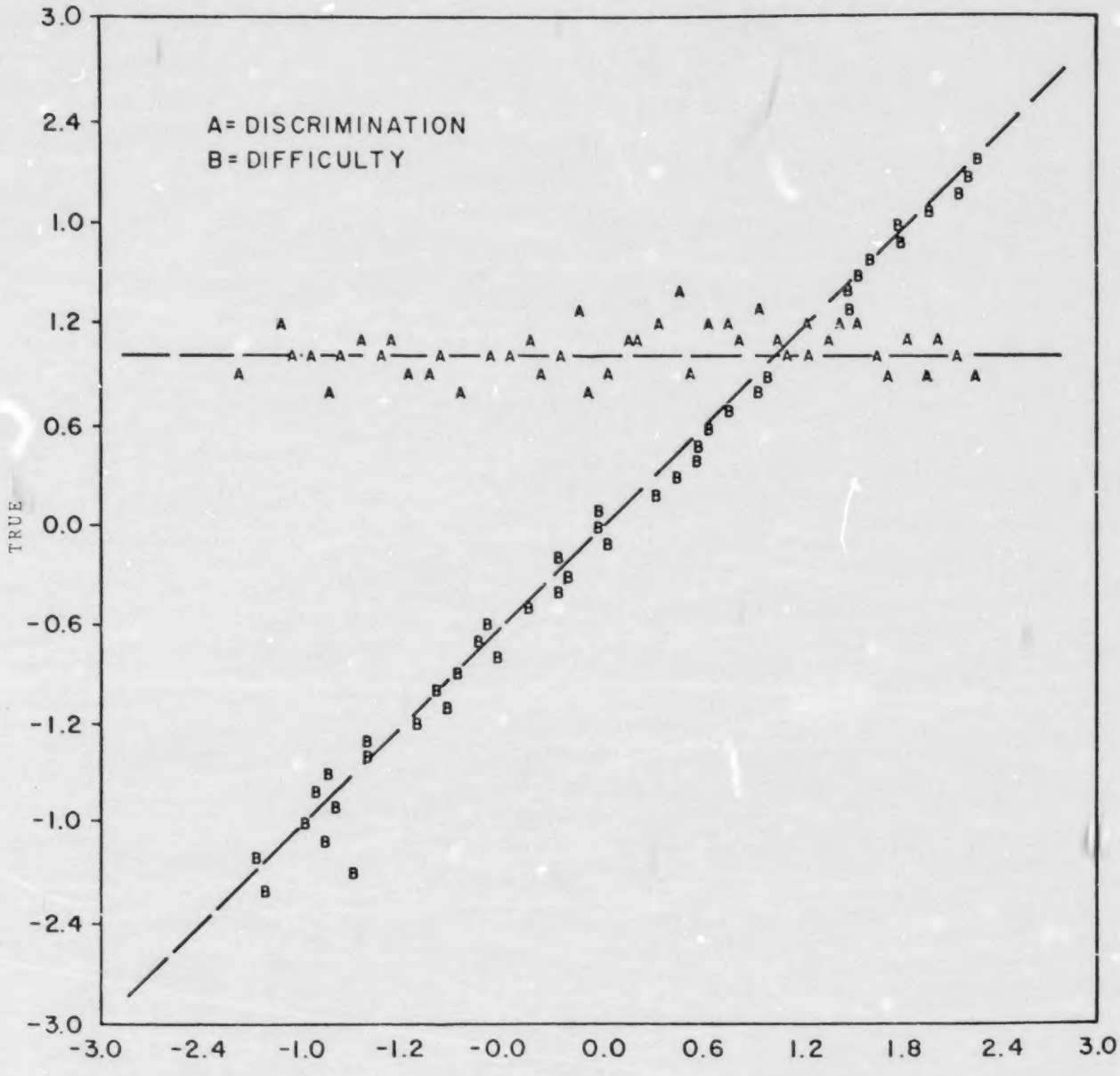
[a] $p < .053$

[b] $p < .004$



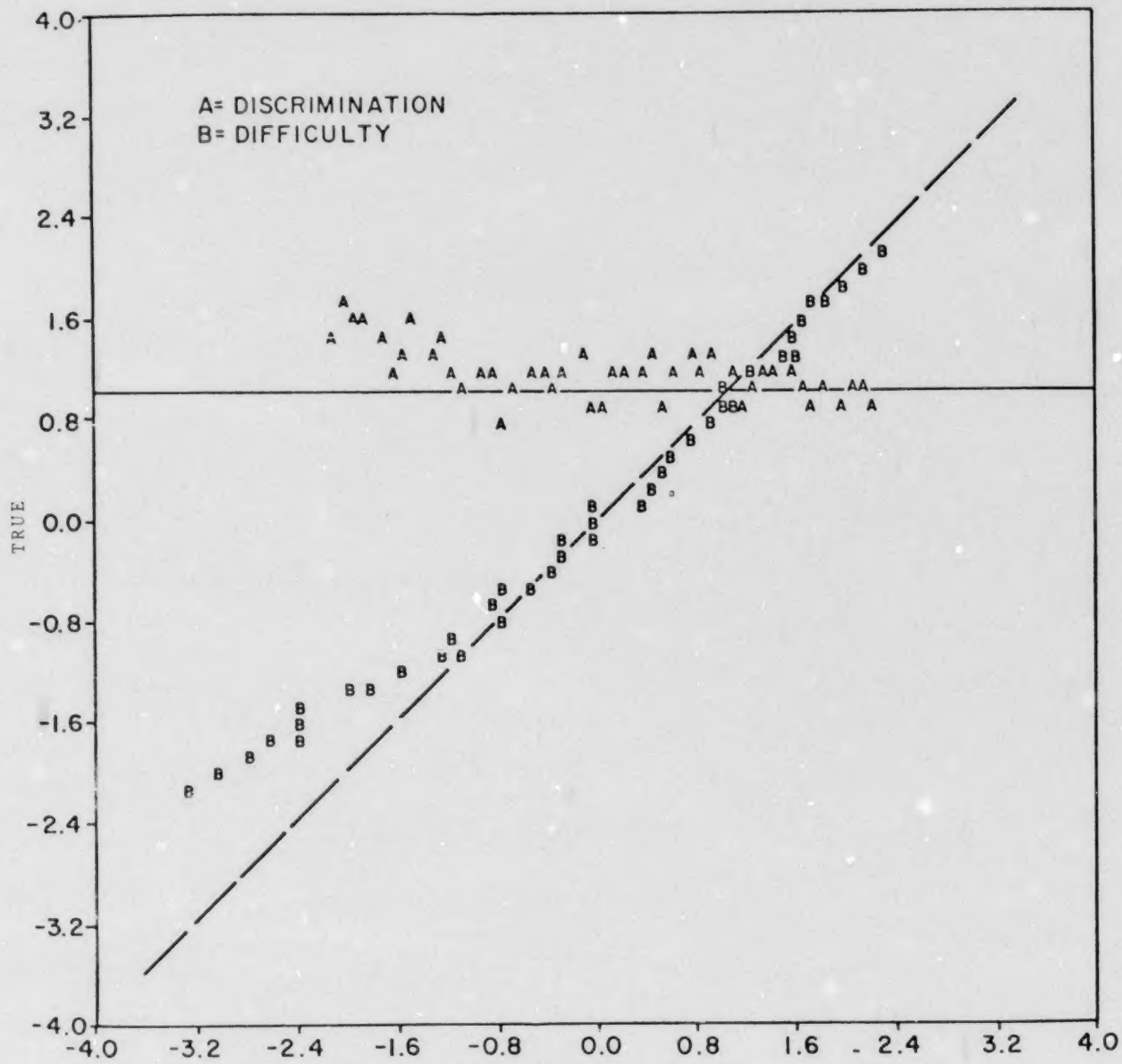
Item parameters from the P_c = .00 analysis of Guessing data (45 Items).

Figure 1



Item parameters from the P_c = .05 analysis of Guessing data (45 items).

Figure 2



Item parameters from the P_c = .10 analysis of Guessing data (45 items).

Figure 3

In both tables we see that the best fit, indicated by the smallest χ^2, occurs at that cutoff point which correctly indicates the percent of random guessers in the sample of simulated examinees. (Consequently with the A.R.R.G. procedure one is able to identify data which is completely free from random guessing as in Table 1 in which case one proceeds with a free response analysis.) The plots of the estimated item parameters (Figures 1-3) from the three analyses of the Guessing data confirm the effect stated in section 4 that over (under)-estimation of the appropriate cutoff point over (under)-estimates (the absolute value of) the item parameters of those items affected by guessing; i.e., the difficult items.

5.2 Data In Situ

The first two of the three instruments analyzed in this paper may be considered subtests of the Survey Test of Educational Achievement (STEA), items for which were selected from the Sequential Tests of Educational Progress (STEP) (Cooperative Testing Service, 1969). These subtests, a reading achievement measure and a mathematics achievement measure, were each formed for this study by combining two of the five subject matter areas which comprise the STEA. The reading subtest is composed of the 25 items which make up the Reading and Mechanics of writing subject matter areas, and the MATH subtest is composed of the twenty-two items which make up the Mathematics computation and Mathematics basic concepts subject matter areas. For the purpose of this study each subtest is considered a separate test administered to a different group of examinees. The fifth grade math subtest and the tenth grade reading subtest were analyzed.

The STEA was given to a very large number of fifth grade and tenth grade students, total N = 39,000, throughout the southern United States. From each population a random sample of size 4000 was selected for analysis relating to the project for which the STEA was developed. From each of these samples a further random selection was

made to reduce the size of each sample for this study to approximately 500.[7]

The third instrument is the 50 item Word Knowledge subtest of Metropolitan Achievement Test. The sample used in this study is a random sample of size 500 taken from the 17,000 fourth graders who participated in the Compensatory Reading Study.[8]

For each test values of P_c from .00 (free response) to .20 in steps of .05 were utilized to obtain the best fitting cutoff point, and for each test the best fit occurred with $P_c = .10$.

The three instruments were analyzed by three models: A free response analysis, an A.R.R.G. analysis, $P_c = .10$, and an item-guessing analysis. In a sense the inclusion of the free response analysis in this section is superfluous: The fit from the A.R.R.G. procedure when applied to free response data will indicate that the data is free from guessing and that all responses should be used in estimation. Therefore, the free response analysis is, when warranted, in-

---

[7] The computer program used in performing the latent trait analyses is a modified version of NORMOJ, Normal Ogive Item Analyser, written for the IBM 360/65 at the University of Chicago Computation Center (Kolakowski and Bock, 1970), and modified by the author.

[8] These data are a part of the Compensatory Reading Project Contract No. OEC-71-3715. Any conclusions are those of the author and are not necessarily endorsed by the U.S. Office of Education.

cluded in the A.R.R.G. procedure.

Table 3 contains the fits produced by the different analyses. In every set of data the improvement in the fit of the model which accrues from the use of the A.R.R.G. procedure is significant. The implications of this for latent trait item analysis are far reaching.

TABLE 3

χ^2 GOODNESS OF FIT

Test	No. of Items	DF	Free Response	DF	A.R.R.G.	DF	Item-Guessing
Word Knowledge	50	398	995.0[c]	398	700.5[b]	353	975.1[c]
Reading	25	198	297.0[a]	198	229.7	173	261.4[a]
Mathematics	22	174	259.3[b]	174	217.5[a]	152	266.2[b]

[a] $p < .01$.
[b] $p < .001$
[c] $p << .001$

The attraction of latent trait models results from their ability to admit measurement on a scale with a well-defined metric which in turn results from the probabilistic assumption concerning the form of the response relation. Under either the free response or item-guessing models, analyses of the

Reading and Mathematics tests result in rejection of this assumption for each instrument as a whole. In this circumstance one recommended procedure for the item analyst is to investigate the individual items in an attempt to determine which items are failing to fit the model. It is suggested that such items be either removed from the instrument or returned to the item constructor for rewording (Lord and Novick, 1968). With each of these measuring instruments, however, the A.R.R.G. analysis reveals that either option may be contraindicated; the error in the item-analysis procedure lies not with the items, but with the failure of either model to adequately remove the effects of random guessing from the analysis. For a test in which significant lack of fit is found during item analysis, the A.R.R.G. procedure results in fewer items being examined and/or eliminated.

Parenthetically we note that the item-guessing analysis fails to converge on some response vectors corresponding to very low ability levels. In the two STEA subtests, twenty subjects in reading and twenty-one subjects in math were inestimable with this analysis, while twelve subjects were lost in the analysis of the word knowledge instrument. This is an example of the result presented by Samejima (1973), indicating that under the item-guessing model, maximum likelihood estimates corresponding to certain response vectors may not be unique or may not even exist at finite values.

In this regard, in the analyses of all three instruments A.R.R.G. failed to produce an ability estimate for only one subject. This subject received credit for only two items out of the twenty-two math items, and attempted every one of these items.[9] Since two out of twenty-two do not quite differ significantly from the chance percentage of 25 per cent[10] ($p = .0606$), we suggest that this subject guessed at most of these items and that measurement of him by this instrument is inappropriate.

---

[9]With respect to guessing STEA examinees are instructed as follows: "If a question seems to be too difficult, make the most careful guess you can'" and "Wrong answers will not be counted against you."

[10]STEA items have four alternatives.

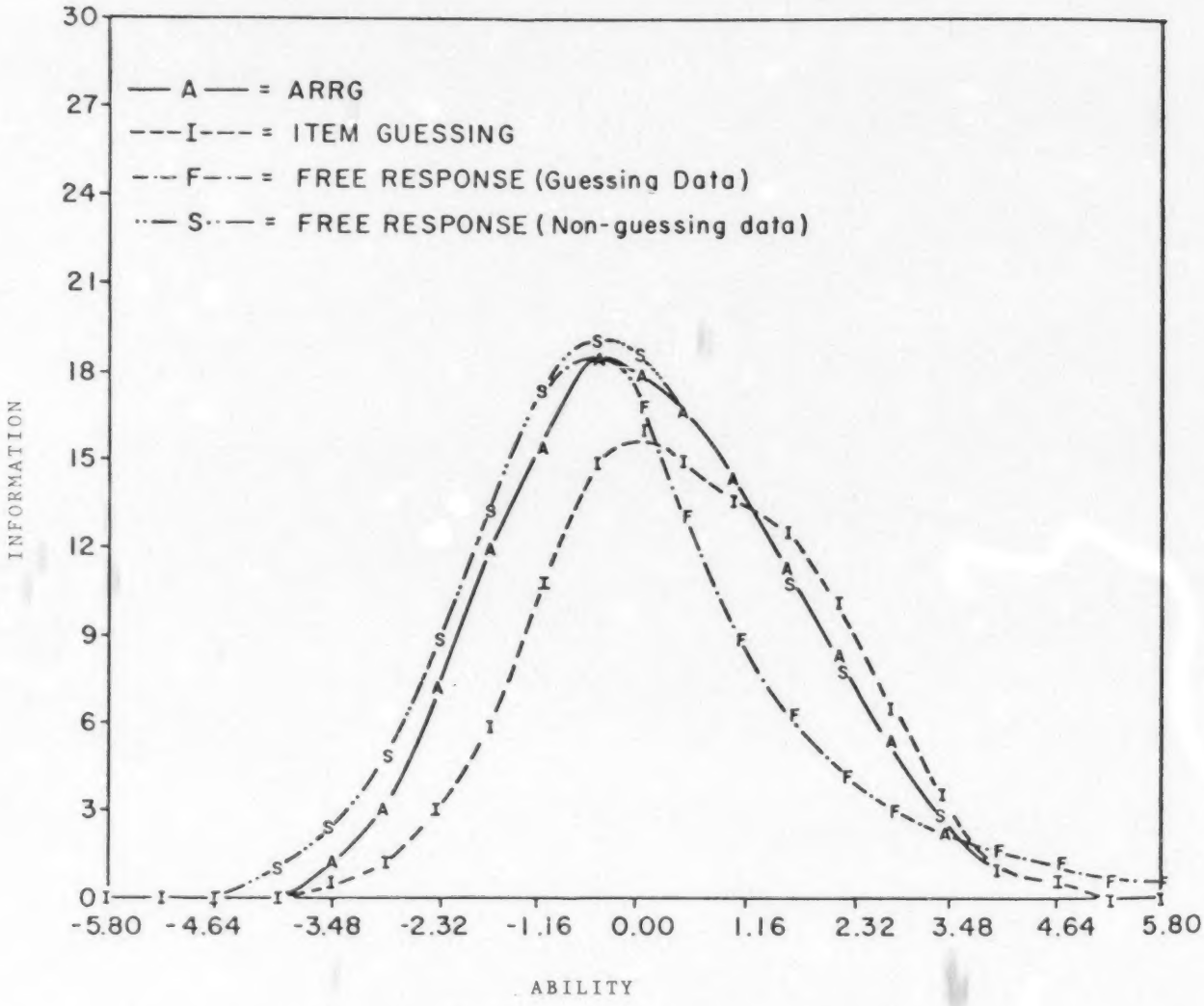
6. The Information Structure

Appropriate comparisons of the information structure as estimated by the different analyses of the Guessing data provide an insightful basis for evaluation of the different models. Consider the loss of precision or information that one might expect to result from random guessing. If we keep item parameters and subject parameters constant, the effect on information of random guessing should be to lower the amount of information concerning estimation of only those abilities in the lower portion of the ability continuum. Generally speaking, only lower ability people have the opportunity to do much random guessing; clearly the farther up the ability continuum, the less the opportunity to guess. The information structure recovered from an analysis should reflect this situation. We will proceed to examine the information recovered by three analyses of simulated item responses contaminated by guessing: the free response analysis (F), the A.R.R.G. analysis (A) and the item-guessing analysis (I).[11]

We will use as a standard the information provided by the free response analysis of the Non-guessing (i.e., free response) data with constant discriminations. Figure 4 contains information curves from such a free response analysis of Non-guessing data(S) and from the three analyses of the Guessing data. Recall that except for the simulated random guessing, the two sets of data were generated using

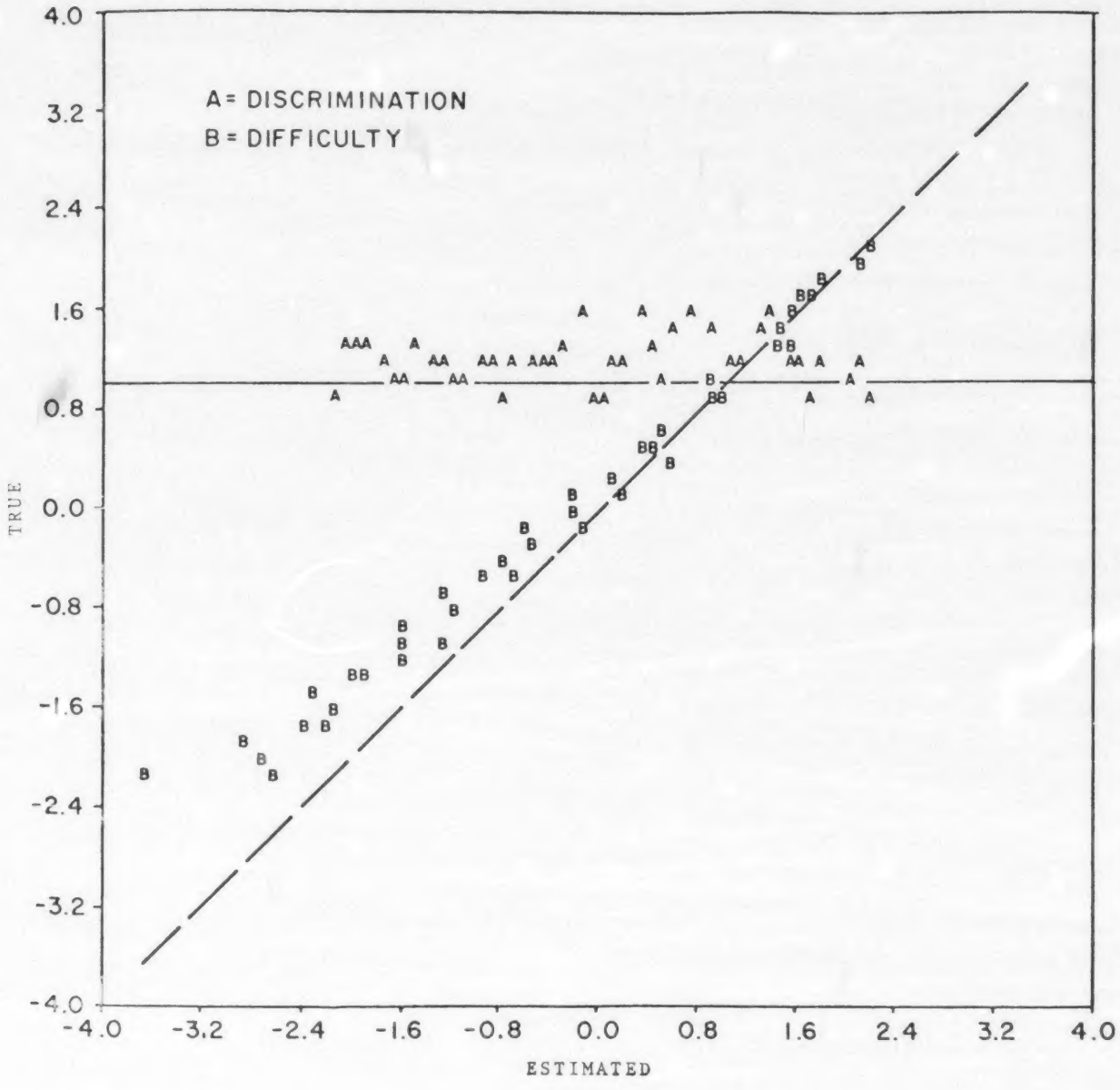
---

[11] See Lord (1968, p. 1014) for a description of the estimation procedure used for the guessing parameter in the 3-parameter item-guessing model.



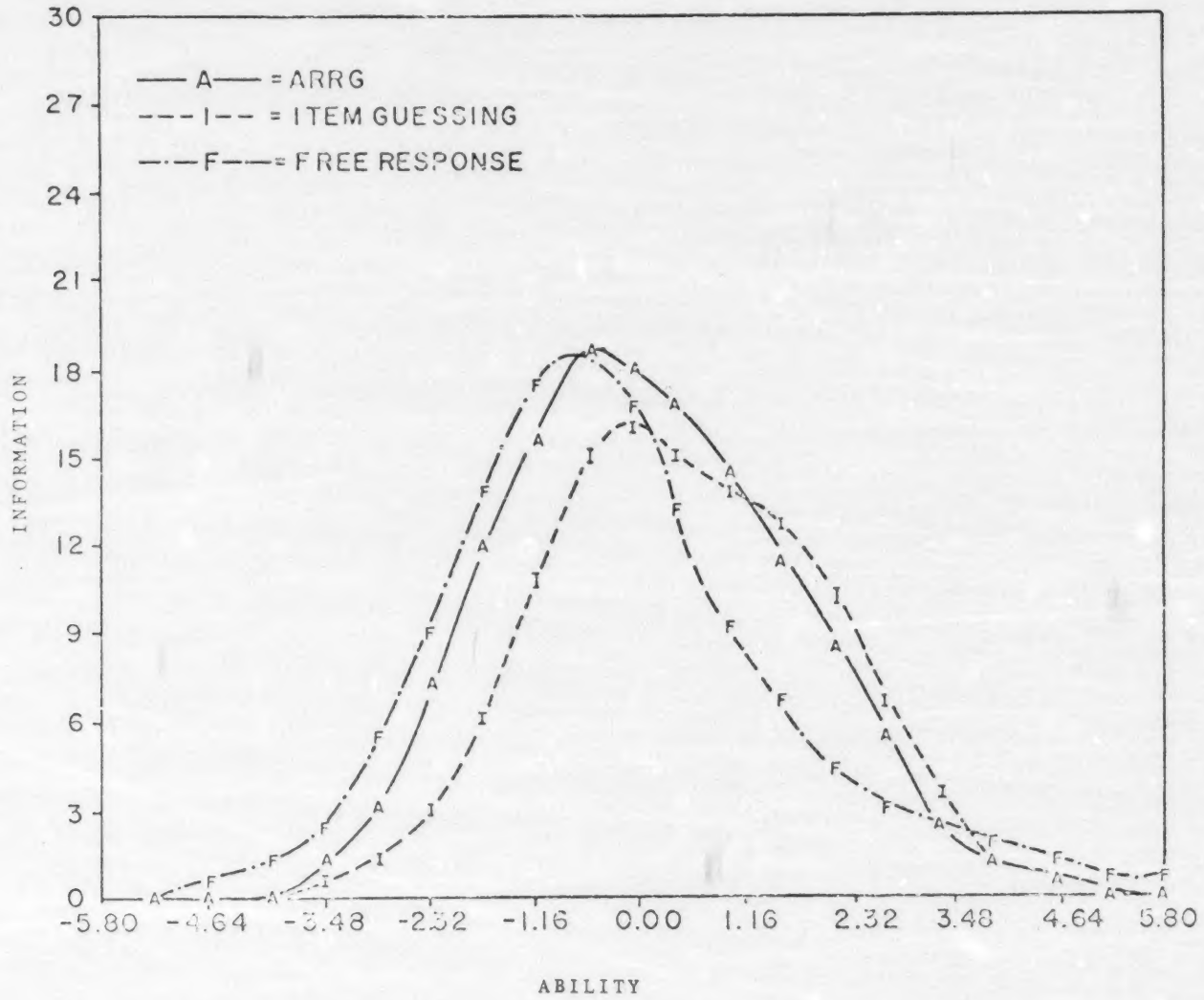
Information curves from analyses of simulated data.

Figure 4



Item parameters from Item-guessing analysis
of Guessing data (45 items).

Figure 5



Information curves from analyses
of simulated Guessing data

Figure 6

the same model parameters and identical sequence of random numbers. Consequently we may use this curve as an approximation to the best one might be able to recover in estimated information for a test with these true item parameters and this distribution of abilities. (In fact, while the S information curve was drawn using the estimated item parameters, the true information curve, which by virtue of our privileged knowledge of the true item parameters is also available to us, is not distinguishable on this scale from the S curve. We note that since every information curve presented in this article was drawn using the estimated item parameters each is in fact an estimated information curve. Of course either information curve is drawn with level of ability as the abscissa so that estimates of ability do not enter in to the determination of these curves.)

Our first comparison is between the A.R.R.G. analysis of Guessing data and our standard, the free response analysis of Non-guessing data. The resulting information curves are as they should be: Guessing has resulted in poorer precision at lower ability levels, and equal precision elsewhere.

Our next comparisons concern the information curve of the free response analysis of guessing data. This curve apparently reports more information concerning subjects in the lower half of the ability range than the A.R.R.G. analysis.[12] A comparison to the S curve shows that the free response analysis of guessing data reports as much information con-

---

[12] The F curve is covered by the S curve in this region.

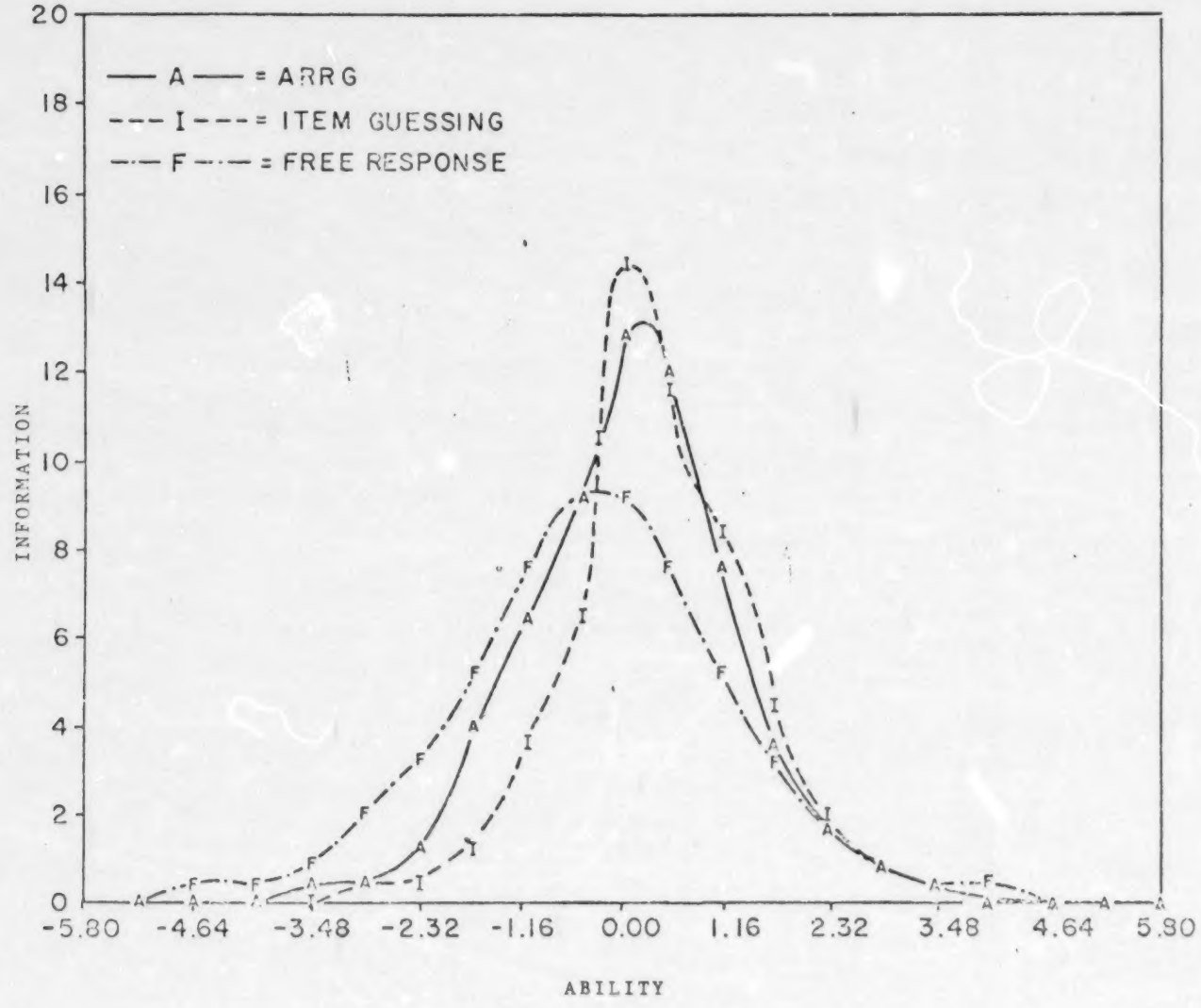
cerning lower ability subjects as the free response analysis of truly free response data. Substantively we know this must be an erroneous conclusion as random guessing must result in a smaller amount of information for levels of ability where such behavior occurs. Proceeding along the ability continuum we observe a significant decrease in the amount of recovered information by this analysis for higher level subjects, subjects with little or no opportunity to reduce information by guessing! The estimated item-parameters of a free response analysis of guessing data provide an explanation (Figure 1); guessing lowers the estimated discriminations of the difficult items, thus lowering the amount of estimated information in the upper half of the ability continuum. Regarding the lower half, guessing doesn't affect estimation of easy items, consequently the estimated information from the free response analysis erroneously reports as much information at these levels as that which would be recovered from truly free response data. The A.R.R.G. analysis in this range omits the component of information resulting from low-ability-subject, high-difficulty-item interactions, thus more accurately representing the amount of information available for estimating these subject parameters.

Our final comparisons concern the information curve of an item-guessing analysis of the simulated guessing data (I). There is a clear improvement in recovered information at the lower ability levels by the A.R.R.G. analysis as contrasted to the item-guessing analysis. The difference between the two may, perhaps, be a general result of the item-guessing analysis' complete failure to take into account individual differences in guessing tendency. More interesting,perhaps, is that apparently the item-guessing analysis of simulated guessing data recovers more information at the upper ability levels than a free response analysis of identically generated--and therefore appropriate for this comparison--free response data. The item-guessing analysis seems to indicate that the introduction of guessing into the data at the lower half of the ability continuum results in an increase in the information at the upper half. Mathematically the reason for this paradox is clear, the item-guessing analysis overestimated the item parameters in the upper portion of ability continuum (see Figure 5). Psychometrically, however, it is difficult to rationalize at any ability level an increase in information as a result of random guessing.

In this figure we've used the simulated guessing data generated by the A.R.R.G. model. When comparing the information structure from the three analyses of data in situ we observe the same general pattern in information curves that we've

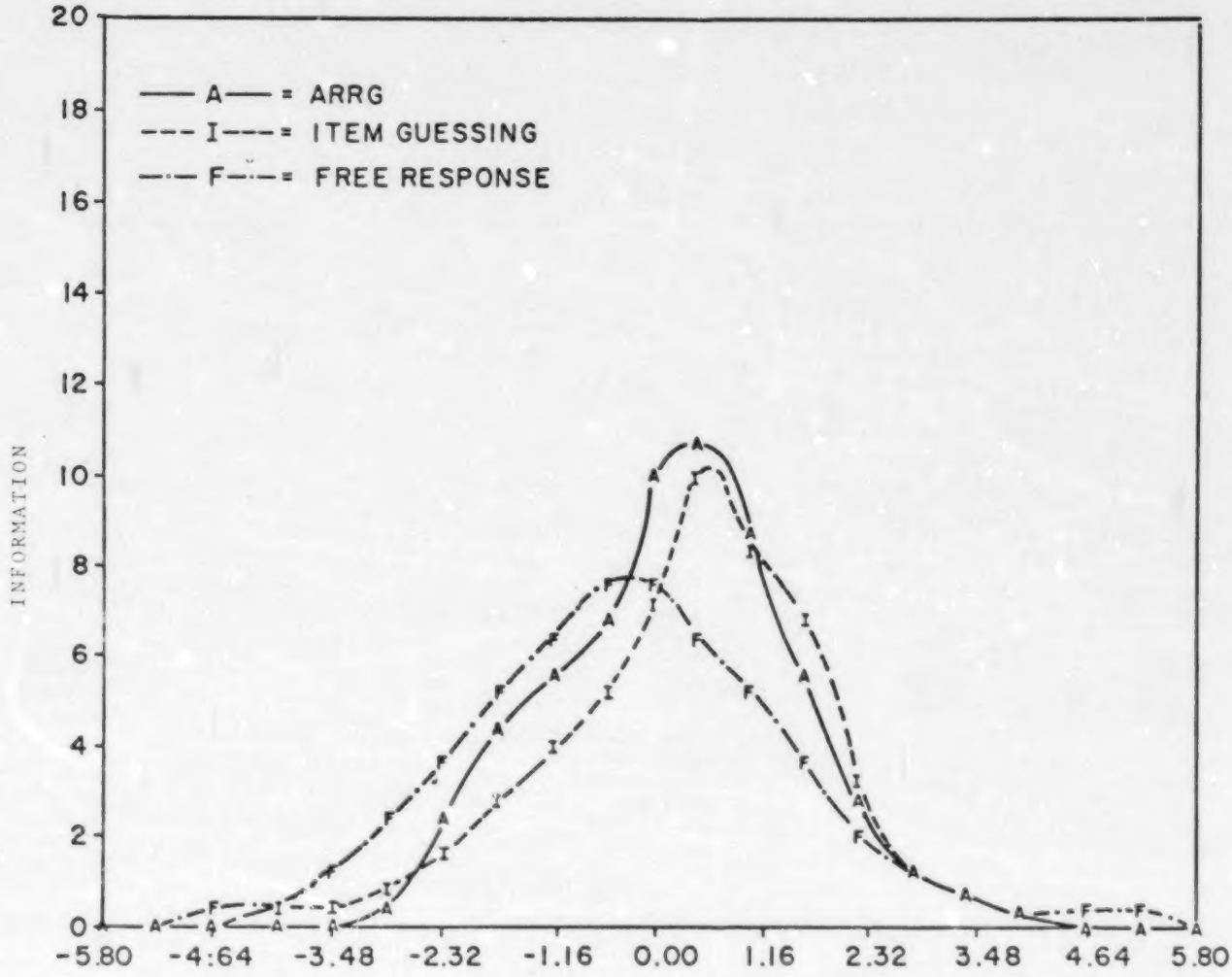
outlined here with simulated data. Of course we don't have an appropriate set of free response data to make comparisons of information curves, but we are able to obtain curves such as Figure 6: Information curves of three analyses, free response, A.R.R.G., and item-guessing, of multiple choice data, data we assume contains random guessing.

Figures 7, 8, and 9 contain the information curves from the different analyses of the three tests described above. A comparison of these figures to Figure 6, reflecting an analysis of data known to be generated by the A.R.R.G. model, immediately indicates the similarity of the structure of all four figures. Each figure is characterized below the mean ability by the ordering: free response, A.R.R.G., item-guessing. At some point below the mean ability the information reported by the free response begins to decline and falls below the other two curves for higher abilities. At approximately the mean ability the three parameter item-guessing analysis begins to report more information than the A.R.R.G. As argued above this is a result of an overestimation of the information resulting from overestimation of the true item parameters by the item-guessing analysis. We feel that these similarities between analyses of data known to be generated by the A.R.R.G. model, and the parallel figures based on analyses of real multiple choice data sets, lend further support for use of the A.R.R.G. model in analyses of data contaminated by guessing.



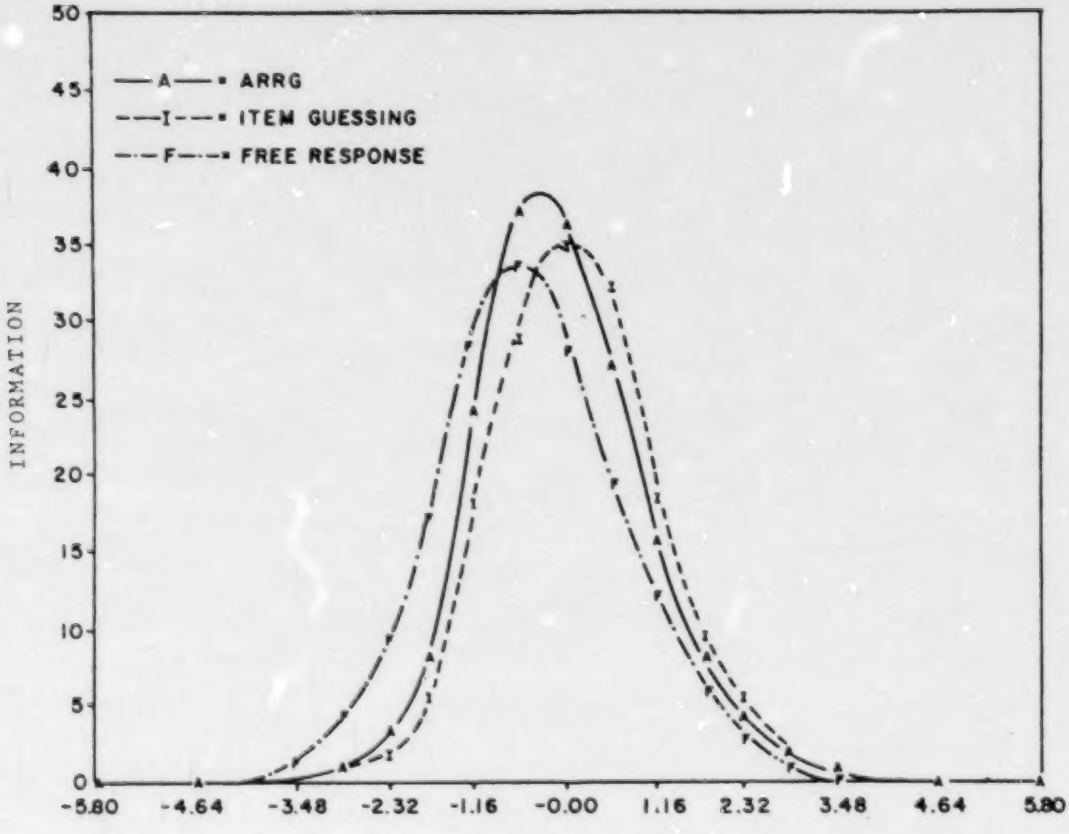
Information curves from analyses of the Reading Test.

Figure 7



Information curves from analyses of the Mathematics Test.

Figure 8



Information curves from analyses of the Word Knowledge test.

Figure 9

We may also compare the different analyses in terms of average information. If the underlying ability, θ, is distributed normally with mean zero and variance one in the population and the estimate of θ is scaled accordingly, the average information is

$$\overline{I(\theta)} = 1/\sqrt{2\pi} \int_{-\infty}^{\infty} I(\theta) \exp(-\theta^2/2)\, d\theta$$

which is readily evaluated by Gauss-Hermite quadrature. For these three tests the average information for each analysis is given in Table 4.

Table 4

Average Information

Test	Free Response	A.R.R.G.	Item-guessing
MAT Word Knowledge	22.68	25.84	24.51
Reading	6.32	7.84	6.63
Mathematics	7.33	9.62	8.47

We see that for each test the A.R.R.G. analysis provides the greatest average information for ability estimation.

REFERENCES

Birnbaum, A. (1968). Some Latent Trait Models and Their Use in Inferring an Examinee's Ability. In Lord, F. M. and Novick, M. Statistical Theories of Mental Test Scores. Reading, Massachusetts: Addison-Wesley.

Bock, R. Darrell (1972). Estimating Item Parameters and Latent Ability when Responses Are Scored in Two or More Nominal Categories. Psychometrika 37, 29-51.

Bock, R. Darrell and Jones, Lyle V. (1968). The Measurement and Prediction of Judgment and Choice. San Francisco, California: Holden-Day.

Bock, R. Darrell, and Lieberman, Marcus (1970). Fitting a Response Model for N Dichotomously Scored Items Psychometrika 35, 179-199.

Cooperative Testing Service (1969). Sequential Tests of Educational Progress (STEP), Version II, Norming Edition. Published by Educational Testing Service Cooperative Testing Division.

Diamond, J. and Evans, W. (1973). The Correction For Guessing. Review of Educational Research 43 (2), 181-191.

Hildebrand, F. B. (1956). Introduction to Numerical Analysis. New York: McGraw-Hill Book Co., Inc.

Kolakowski, Donald and Bock, R. D. (1970). A Fortran IV Program for Maximum Likelihood Item Analysis and Test Scoring: Normal Ogive Model. Statistical Laboratory Research Memorandum No.12. Department of Education, University of Chicago.

Lord, Frederic M. (1968). An Analysis of the Verbal Scholastic Aptitude Test Using Birnbaum's Three Parameter Logistic Model. Educational and Psychological Measurement 28, 989-1020.

Lord, Frederic M. and Novick, M. (1968). Statistical Theories of Mental Test Scores. Reading, Massachusetts: Addison-Wesley.

Panchapakesan, Nargis (1969). The Simple Logistic Model and Mental Measurement. Unpublished Ph.D. Dissertation, University of Chicago.

Samejima, Fumiko (1972). A General Model for Free Response Data. Psychometric Monograph No. 18.

Samejima, Fumiko (1973). A Comment on Birnbaum's Three Parameter Model in the Latent Trait Theory. Psychometrika 38, 221-233.

Waller, Michael I. (1974). Estimating Guessing Tendency. Paper presented at the Annual Convention of the Psychometric Society, San Francisco, California, March 29, 1974.